Response to Reviewer 2

General comments:

This manuscript calculated the apparent turnover time of top and subsoil SOC on a global scale. The major outcome from this work is very useful and timely needed for soil biogeochemistry and carbon cycle modeling communities. The comprehensive data inputs authors used, random forest based geospatial predictive mapping and in-depth uncertainty analysis guaranteed the quality of the produced τ map. Overall, this is a solid and interesting work and I would like to recommend it for acceptance after some technical revisions.

AC: Thank you for your kind words. We are delighted that you acknowledge the importance of our work and thank you for your positive conclusion on our manuscript.

Specific comments:

The description of quantile regression forest is not very straightforward. Can authors double check and revise this part?

AC: Thank you for your suggestion. The quantile regression forests (QRF) model estimates the quantiles of the conditional distribution of the target variable at prediction points. For example, the 0.05 and 0.95 quantiles are typically derived to represent the lower and upper limits of a symmetric 90% prediction interval. QRF first constructs a RF model in the usual way, by developing multiple decision trees that use subsets of the training data, whereby the prediction of each tree equals the average of the observations in the end node of the tree in which the prediction point sits. The RF prediction is the average of all tree predictions. Since averaging is a linear process, the RF prediction boils down to a weighted sum over the n observations of the response variable:

$$\hat{y}(x) = \sum_{i=1}^{n} w_i(x) \cdot y_i$$

In QRF, the weights $w_i(x)$ are used to estimate the cumulative distribution $F(y|x) = P(Y \leq y|x)$ of the response variable Y, given the covariate data x, as follows:

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_i(x) \cdot 1_{\{y_i \leq y\}}$$

where $1_{\{y_i \leq y\}}$ is the indicator function (i.e., it is 1 if the condition is true and 0 otherwise). Any quantile q of the distribution can then be derived by iterating towards the value of y for which $\hat{F}(y|x) = q$ (Meinshausen, 2006).

We added the above in Section 2.4.2 of the revised manuscript to make the description of QRF clearer.

References:

Meinshausen, N.: Quantile Regression Forests, Journal of Machine Learning Research, 7, 983–999, 2006.

There lacks certain discussion of the topographical effect on τ. Since the topography largely impacts

τ in tundra covered regions (Fig. S23, S24), can authors discuss this finding?

AC: We appreciate the reviewer's insightful suggestion. In the main text, we extracted top six important covariates to analyze their directional effects on topsoil and subsoil τ, according to the feature importance results by random forest model at the global scale (Fig. S23a, S24a). Although we hope to focus on describing these six key factors' effects on τ at the global scale in Section 3.3.1, we acknowledge that the effects of topographical factors should also be explained here. In the revised manuscript, at the end of Section 3.3.1, we added a paragraph to discuss it, and highlighted the large impacts of topography on τ in tundra areas and its potential reasons.

Technical corrections:

I suggest authors double check and add units to all variables in your equations.

AC: Thanks for pointing this out. We added the units of SOC content after showing Eq. 1.

Line 89: "collected form" shall be "collected from"

AC: Corrected. Thanks.

Line 99 - 100: In equation 1, SOC_Du-Dl is not defined.

AC: Thanks. We added the definition of this variable in Eq. 1 in the revised manuscript.

Line 100: "Were" shall be "Where"

AC: Corrected. Thanks.

Line 103: "fit layers observations at different depth intervals" can be simplified "fit observations along depth"?

AC: Corrected. Thanks.

Line 105: "the SOCS at two layers for" shall be more specific "the SOCS at top- and subsoil layers for"

AC: Corrected. Thanks.

Line 112: "The flux of carbon at a certain soil layer needs to be obtained through" can be revised "Carbon influx at each soil layer comes from"

AC: Corrected. Thanks.

Line 114: "The annual NPP (kg C m-2 yr-1) produced by the moderate-resolution imaging spectroradiometer (MODIS)" means the MOD17 products? Please add the specific version of the MOD17 product and the url where the author downloaded this data.

AC: Thanks for the suggestion. We added the specific version of this data product and the URL in the revised manuscript. We also provided the reference (Running and Zhao, 2019) (in the main text and Table S1) related to this dataset when describing it.

Reference:

Running, S. and Zhao, M.: MOD17A3HGF MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global

500 m SIN Grid V006, https://doi.org/10.5067/MODIS/MOD17A3HGF.006, 2019.

Line 126: "total amount of roots" shall be "total root biomass"?

AC: Corrected. Thanks.

Line 132: " the root distribution" shall be " the root biomass distribution"?

AC: Corrected. Thanks.

Line 144: " for each soil sample site, root profile observations within the same terrestrial ecoregions (Dinerstein et al., 2017) and the same soil type (FAO–Unesco, 1990) as that of the soil sample were selected. The corresponding mean of those selected root observations for each soil sample location were finally collected (Figs. S8 and S9)" can be simplified "we apply the arithmetic mean of fr from root profile observations within the same terrestrial ecoregions (Dinerstein et al., 2017) and soil type (FAO–Unesco, 1990) as soil sample."

AC: We revised this sentence following your suggestion. Thanks.

Line 189: "The partial correlation of each influencing factor was calculated while controlling other factors" at which level? Mean or median?

AC: We appreciate the reviewer's request for clarification. The partial correlation of each influencing factor was calculated while controlling for other factors at the mean level using Pearson's partial correlation analysis. This approach estimates the linear relationship between τ and each environmental factor while holding other variables constant, based on their mean values. To improve clarity, we have added "at the mean level" in this sentence for improving clarity following your suggestion.

Line 201: "and this division was performed on each biome data to ensure that the ten split sets can keep a balance among biomes". If my understanding is correct, do you mean "and samples of each biome in each subset has the same proportion as the whole dataset"?

AC: Yes. We revised this sentence following your suggestion. Thanks.

Line 209: "by replacing observations by indicator transforms". Not sure I understand "indicator transforms". Do you mean replacing the actual observation values by a tag of category? If so I'm not sure why this is needed. Can you provide a simple example?

AC: We appreciate the reviewer's request for clarification. In QRF, the term indicator transforms refers to the approach used to estimate the conditional quantiles of the target variable, rather than categorizing observations. Please see our response to the first comment. The term "indicator transforms" referred to the use of the indicator function. To improve clarity, we have revised related contents in Section 2.4.2 to make the description of QRF method clearer.

Line 213: "has been also" shall be "has also been"

AC: Corrected. Thanks.

Line 215: If my understanding is correct, it's nice to consider error propagation and produce a dataset considering this uncertainty information to train quantile regression forests. But in your writing this information is not explicitly conveyed so I feel a bit confused when reading this part. It will be better if authors can explicitly tell readers this information above this paragraph.

AC: We consider the uncertainty mainly comes from two sources of error. The first is the model error, which refers to covariates that do not fully explain the variations in the target variable (i.e., top- and subsoil τ in this study) and error in the estimation of the model parameters. The second is the measurement error, which represents the difference between the actual and recorded value that related to the calculation of the target variable (i.e., the input variables related to τ calculation) (van der Westhuizen et al., 2022; Heuvelink and Webster, 2023). Therefore, the first two paragraphs in Section 2.4.2 were written to describe how did we considered these two sources of error. To better let the reader understand the ideas of uncertainty calculation, we added some content at the beginning of the first paragraph in Section 2.4.2, to make a linkage between the first source (using QRF to calculate) and the second source (using error propagation to calculate the uncertainty of inputs) of error.

References:

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D.P., Poggio, L., 2022. Measurement error-filtered machine learning in digital soil mapping. Spatial Statistics 47, 100572. https://doi.org/10.1016/j.spasta.2021.100572

Heuvelink, G.B.M., Webster, R., 2023. Uncertainty assessment of spatial soil information, in: Encyclopedia of Soils in the Environment. Elsevier, pp. 671–683. https://doi.org/10.1016/B978-0-12-822974-3.00174-9

Line 250: Unfinished sentence "The shaded grey area represents the"

AC: Thank you for your careful review. This is a sentence from leftovers in the working version in the original manuscript. We removed it in our revised version.

Figure 6: is interesting. I guess the y axis title "MAT/MAP effect" means the partial regression coefficient between MAT/MAP and τ? Please add a description in figure caption.

AC: Yes, you are correct. We added a sentence in the figure caption as you suggested.

Figure 6: Another question. I know to completely decorrelate climate and edaphic variables is extremely tough, but I would suggest authors provide a correlation matrix to visualize and identify potential issues with "multicollinearity" between 2 climate and 4 edaphic variables.

AC: Here we plot a heatmap figure to show the pairwise correlation between those 6 covariates. It shows that most of them have a relatively low (< 0.4) Pearson correlation coefficient to each other, except pH and MAP. Therefore, we think multicollinearity issue is not significant here.

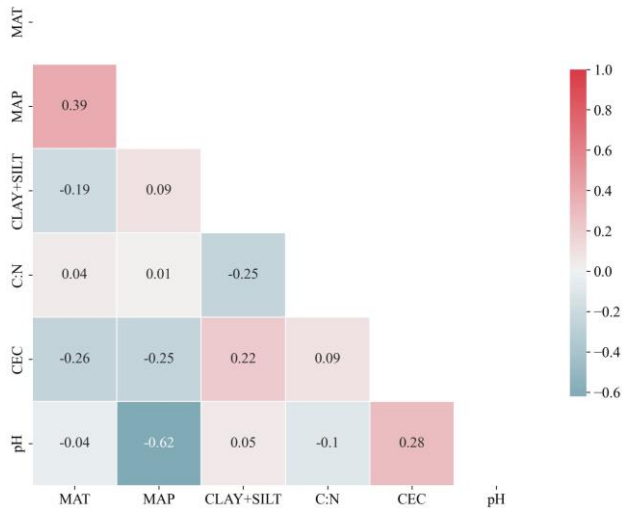

Figure | Correlation between two climate and four edaphic variables using Pearson correlation analysis.

Figure 6d shows different responses for topsoil and subsoil. As high CEC enhances adsorption, more nutrients would be available for micro-organism and more mineral-organic compounds are formed, which tends to increase sensitivity of SOCS to temperature. But meanwhile, plants are more productive and may have higher plant carbon stock, thus might increase sensitivity of NPP to temperature. The results might be the combination of both effects.

AC: Thank you for this insightful comment. We agree with your idea and added this discussion in the revised manuscript (Section 3.2.2).

Figure 7c: What is the unit of variable importance (%)?

AC: Yes, the unit is %. We revised the figure to add this unit.

Line 405: "biogeochemical simulations by ESMs, and will be useful to improve" can be simplified to "ESM simulations and improve"

AC: Corrected. Thanks.

Line 407: shall mention CMIP6 outputs from which experiment?

AC: We analyzed historical simulations outputs of selected ESMs from CMIP6. The historical scenario simulations (also known as the 20th-century simulations) for CMIP6 were carried out for the period from the start of the industrial revolution to near present: 1850–2015. We added a sentence in Section 3.4 to mention it.

From Fig. S17, it seems like most of the wetland samples are collected from the tropics. Would you justify the performance of wetland τ the arctic region?

AC: We appreciate the reviewer's observation regarding the distribution of wetland samples. It is true that a significant portion of the wetland data in our study originates from tropical regions, which may influence the model's ability to accurately capture τ variations in arctic wetlands. However, we have taken several steps to mitigate potential biases and ensure reasonable performance in the arctic region. First, our dataset does include arctic wetland observations from sources such as the Northern

Circumpolar Soil Carbon Database (Hugelius et al., 2013) and other high-latitude soil profiles from WoSIS database. These data contribute to model training and validation, particularly for permafrost-affected regions. Second, the model does not rely solely on the spatial distribution of training samples but also incorporates key environmental variables (e.g., temperature, precipitation, soil physio-chemical properties, and topography) that distinguish arctic from tropical wetlands. This helps the model extrapolate τ values in arctic regions based on known biogeochemical relationships rather than direct spatial proximity. Finally, we extracted our modelling result in the arctic wetland region, and showed the validation accuracy in topsoil and subsoil as follow. Generally, the we obtained an acceptable accuracy on validation set, which justifies the performance of τ predictions in this region.

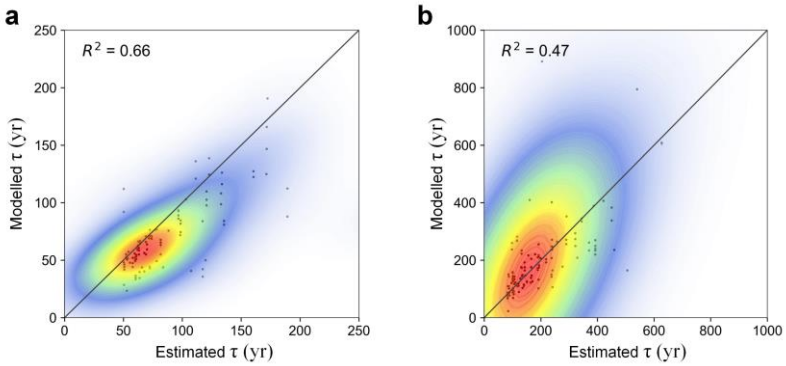

Figure | Validation plots for predictions of soil organic carbon turnover time (τ, yr) in top- (a) and subsoil (b) layer.

References:

Hugelius, G., Tarnocai, C., Broll, G., Canadell, J. G., Kuhry, P., and Swanson, D. K.: The Northern Circumpolar Soil Carbon Database: spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions, Earth System Science Data, 5, 3–13, https://doi.org/10.5194/essd-5-3-2013, 2013.